Mammographic Breast Density Assessment: A Methods Study

Lusine Yaghjyan¹, Susan M. Pinney¹*, Mary C. Mahoney², Arthur R. Morton³, Jeanette Buckholz⁴

¹ University of Cincinnati College of Medicine, Department of Environmental Health, Cincinnati, OH, USA; ² University of Cincinnati College of Medicine, Department of Radiology, Cincinnati, OH, USA; ³ University of Cincinnati University Hospital, Department of Radiology, Cincinnati, OH, USA; ⁴ University of Cincinnati College of Medicine, Department of Internal Medicine, Cincinnati, OH, USA

Received: April 12, 2010 / Accepted: March 24, 2011

Abstract

This pilot study was designed to find the best way to characterize mammographic breast density for future epidemiologic studies. Fifty mammograms (24 high, 6 low, and 20 intermediate density mammograms, as defined by the original radiologist reading) were selected from women participating in a community medical surveillance program. Women were required to be Caucasian, 40-85 years old and have a Body Mass Index ≤30 at the time of the mammogram; and to have no history of breast cancer. Mammograms were selected by stratified random sampling on the year of mammogram and degree of density. The agreement between breast density category assignment using different methods was assessed using the kappa statistic. Substantial to perfect agreement was observed between density categories as determined from two radiologists’ readings (kappa statistic [K]=0.73). Perfect agreement was noted between density of high vs. intermediate (K=0.81) and high vs. low density combinations (K=0.86). When grouped by the year of the mammogram, substantial agreement was observed for all year intervals (K=0.77, 0.69, and 0.70). The agreement between density group assignment using information from digital images and using codes applied to the radiologist reading showed fair to moderate agreement (K 0.19-0.36). Using the radiologist reading for characterization of breast density from earlier conventional mammography films is better than using density quantification from digitized mammograms.

Key words: Epidemiologic methods, breast density, mammography.

Introduction

Mammographic breast density is a strong predictor of breast cancer (Vachon et al., 2007, Harvey & Bovbjerg, 2004, Boy et al., 2006, Haiman et al., 2002, Tamimi et al., 2007, Ursin et al., 2003, Ginsburg et al., 2008). Light (non-radiolucent) areas on the mammogram represent the fibrous and glandular tissues in the breast, whereas, the dark (radiolucent) areas are primarily fat. In clinical practice and research, breast density has been traditionally classified using the Wolfe’s classification with four breast parenchymal patterns (N1, P1, P2, and DY) or the six category classification (SCC) system based on the visual estimation of the percent density (1 [0%], 2 [1–10%], 3 [11–24%], 4 [25–49%], 5 [50–74%], 6 [75–100%]) (Brisson et al., 2003, Petroudi et al., 2003, Highnam et al., 2007, Jeffreys et al., 2003, Petroiu et al., 2003, Highnam et al., 2007, Lee-Han et al., 1995, Jamal et al., 2006). The American College of Radiology proposed a modified version of the Wolfe classification with four breast parenchymal patterns (N1, P1, P2, and DY) or the six category classification (SCC) system based on the visual estimation of the percent density (1 [0%], 2 [1–10%], 3 [11–24%], 4 [25–49%], 5 [50–74%], 6 [75–100%]) (Brisson et al., 2003, Petroudi et al., 2003, Highnam et al., 2007, Jeffreys et al., 2003, Petroiu et al., 2003, Highnam et al., 2007, Lee-Han et al., 1995, Jamal et al., 2006). The American College of Radiology proposed a modified version of the Wolfe classification known as the Breast Imaging-Reporting and Data System (BI-RADS) classification (1-predominantly fat; 2-fat with some fibroglandular tissue; 3-heterogeneously dense; 4-extremely dense) (American College of Radiology, 1998). In addition, several computer algorithms have been developed and used to estimate breast density (Zhou et al., 2001, Muhimmah et al., 2006).

In planning a retrospective study of breast density determinants in a large cohort of women (Fernald Medical Monitoring Program [FMMP]), we conducted a preliminary study to find the best way to characterize mammographic breast density. There were two potential approaches. We could assign a level of density for each mammogram using the radiologist's report. Alternatively, we could digitize the mammograms and evaluate breast...
density using computer-assisted techniques.

A few studies reported a good agreement between breast density estimation by a radiologist and that from the digitized mammogram films (Jeffreys et al., 2008). However, the breast density estimation method used in previous studies differed from the more detailed approach proposed for breast density estimation from the FMMP mammograms. This preliminary study aimed 1) to determine whether the classification of participants into density categories using information coded from the mammography report is as informative as using computer-aided density estimation from digitized mammograms; and 2) to find how FMMP mammography codes could be used to improve definition of low and high density for the planned retrospective study.

**Materials and Methods**

**Mammogram Selection**

Mammograms used in the present study were available from the FMMP mammography database. The FMMP was a community-based medical surveillance program for residents living within five miles of the perimeter of the former uranium processing plant in Fernald, Ohio. All mammograms were read by Board-certified radiologists. Many mammography studies were performed prior to 1995 and thus the BI-RADS classification was not included in those mammography reports. Mammographic reports were assigned codes from a set of over 230 standardized mammography codes that represented the inclusion of specific descriptors in the radiologist’s report. For example, the code 704 (Breasts high in density/markedly dense) was assigned to the following description “The breast tissue is markedly dense…” The text of the report and the codes were entered into an extensive database.

For inclusion in the methods study, women were required to be Caucasian (99% of FMMP population), 40-85 years old and have a Body Mass Index (BMI) ≤30 at the time of the mammogram of interest, to never have had breast cancer, and to be eligible for the main study. Using the mammogram codes assigned to the radiologist’s report, all eligible women were classified into one of the three breast density categories: low, intermediate, and high. Women in high density category were required to have at least one mammogram with the high density code. The intermediate density group consisted of women who never had the high density code on any mammogram, but had one or more intermediate density codes on at least one mammogram. Women with low density were required to have never had any of the high or intermediate density codes. The selection of mammograms was restricted to those which were done as a part of the FMMP annual examinations (97.8%). As the radiologist (MCM) designated to do the second reading for this study was the original reader for some of the FMMP mammograms, those mammograms were excluded from the study (3.5%).

Only a small amount of funding was available and the study was restricted to 50 mammograms. We a priori decided to select a larger numbers of high and intermediate density films (24 and 20, respectively) and a smaller number of low density films (6 films). We also stratified our sampling on the year of mammogram to account for changes in technology over time (Haus, 2002). The mediolateral oblique (MLO) views of 50 mammograms were evaluated in two phases of the study: 1) the second radiologist assessment of breast density and 2) quantification of breast density from the digital image. The breast density assignment from two phases was compared with the level of density from the original mammography report (Figure 1).
Inter-rater Reliability Assessment

All 50 mammograms were re-read by a second Board-certified radiologist from the Department of Radiology (MCM) at the University of Cincinnati (UC). The radiologist reported findings in a standardized form that included both a direct question about the degree of breast density and a list of different mammogram descriptors (linked to the FMMP mammography codes). The radiologist provided a Yes/No response to each of the descriptors. The second radiologist estimated density for right and left breast separately.

For comparison of the two radiologists’ readings, we combined information from both breasts to be consistent with the coding of the original mammogram report. The overall density for the mammogram corresponded to the highest degree of density assigned by the second radiologist to either of the breasts. We compared the original density category and 1) the density level directly assigned by the second radiologist (2nd-direct Q) and 2) the density level corresponding to the descriptors selected by the second radiologist (2nd-descriptors). The inter-rater reliability analysis was first conducted for all 50 mammograms. For statistical analysis, we created subsets of two density categories at a time (high and intermediate, high and low, low and intermediate), and subsets of mammograms by calendar year (≤1995, 1996-2000, >2000).

Quantification of Density from Digital Images

All mammograms were scanned with a standard equipment (DiagnosticPRO Advantage, VIDAR Systems Corporation, VA, USA) available through the UC Department of Radiology, using one of four resolutions (1st: 870 microns/8 bit precision; 2nd: 670 microns/8 bit precision; 3rd: 220 microns/12 bit precision; and 4th: 170 microns/12 bit precision, respectively). Using a single resolution would result in loss of important information as films used in the study were done over a 16 year period and had different technical specifications. Manipulation of the digital images and density estimation were performed using OSIRIS 4 Software (Geneva, Switzerland).

First, the medical physicist (ARM) manually excluded the pectoral muscle area on the digital image; then, the rest of the breast area was divided into three irregularly shaped areas (lower, middle, and upper triads) of approximately equal size (based on the number of pixels). Density within each of the triads was characterized by Pixel Average Grayness (PAG)—an average of pixel grayness levels weighted by the number of pixels.

Statistical analysis of digital images aimed to find a single density estimate for each mammogram that could be used for assignment to a density category. To account for the potential effect of differences in resolution, PAG values for the triads were normalized to resolution 1 (the largest group). The maximum PAG value across all triads for the given mammogram (both breasts) was used as a critical density estimate. The critical density estimates for all 50 mammograms were ranked from low to high, and the mammograms were assigned density categories based on their ranks and according to the distribution of the original density categories (6 low, 20 intermediate, and 24 high). The density group assignment from the digital image was compared against the original breast density category (Figure 1). To explore the effects of resolution adjustment on the analysis results, a restricted analysis was conducted on the subset of mammograms scanned with resolution 1 (baseline group, PAG values never adjusted) and on the subset of mammograms scanned with resolution 3 (using unadjusted PAG values). An additional analysis was conducted on the subset of 39 mammograms that had a perfect agreement between the two radiologists’ readings.

Statistical Analysis

The data were analyzed using SAS (version 9.2, SAS Institute, Cary, NC). The agreement from all comparisons was assessed with the kappa statistic (simple or weighted) (< 0 - less than chance agreement; 0.01–0.20 - slight agreement; 0.21–0.40 - fair agreement; 0.41–0.60 - moderate agreement; 0.61–0.80 - substantial agreement; 0.81–0.99 - almost perfect agreement) (Viera, 2005). PAG values were adjusted for the 1st resolution.

Table 1. Distribution of participant characteristics by density category.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>High (N=24)</th>
<th>Intermediate (N=20)</th>
<th>Low (N=6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at the time of mammogram (Mean (SD))</td>
<td>48.4 (7.3)</td>
<td>55.2 (11.3)</td>
<td>58.3 (8.3)</td>
</tr>
<tr>
<td>BMI at the time of mammogram (Mean (SD))</td>
<td>24.0 (3.3)</td>
<td>25.2 (2.9)</td>
<td>22.4 (1.9)</td>
</tr>
<tr>
<td>Premenopausal women, N (%)</td>
<td>20 (83.0 %)</td>
<td>3 (50.0 %)</td>
<td>11 (55.0 %)</td>
</tr>
<tr>
<td>Year of mammogram, N (%)</td>
<td>≤1995 8 (33.3 %)</td>
<td>5 (25.0 %)</td>
<td>2 (33.3 %)</td>
</tr>
<tr>
<td></td>
<td>1996-2000 8 (33.3 %)</td>
<td>8 (40.0 %)</td>
<td>2 (33.3 %)</td>
</tr>
<tr>
<td></td>
<td>&gt;2000 8 (33.3 %)</td>
<td>7 (35.0 %)</td>
<td>2 (33.3 %)</td>
</tr>
</tbody>
</table>
using Analysis of Variance (ANOVA) for the nested hierarchical design.

The study was conducted under the Institutional Review Board approval from the University of Cincinnati (last approved on January 14, 2009).

**Results**

The distribution of participant characteristics by density category is presented in Table 1. The mean age at the mammogram date for the original high density group (48.4 years) differed significantly from both the intermediate (55.2 years, \( p=0.03 \)) and the low density (58.3 years, \( p=0.01 \)) categories (Table 2). The high density group contained significantly larger proportion (\( p=0.01 \)) of premenopausal women (20/24, or 83%). The mean BMI (kg/m\(^2\)) for the low density group (22.4) differed significantly from BMI for the intermediate density group (25.2, \( t\)-test \( p=0.04 \)).

**Inter-rater Reliability Assessment**

Using the direct question on overall density (2nd direct Q), the second radiologist rated 23 mammograms as high \( [46.00\%] \), 15 mammograms as intermediate \( [30\%] \), and 12 mammograms as low density films \( [24\%] \). When using the information from the descriptors (2nd descriptors), the density category distribution was as follows: 25 \( [50\%] \) high, 5 \( [10\%] \) intermediate, and 20 \( [40\%] \) low density films.

Substantial agreement (weighted kappa statistic \( [wK] \) 0.73) was observed between the original density category assignment and the overall density assignment from the second radiologist reading (Table 2). Almost perfect agreement was noted between two radiologists for subsets of high vs. intermediate and high vs. low density groups (\( K=0.86 \) and \( 0.81 \), respectively).

<table>
<thead>
<tr>
<th>Study group/subset ( ^a )</th>
<th>Original density category vs. 2nd-Direct ( ^b )</th>
<th>Original density category vs. 2nd-Descriptors ( ^c )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Exact Agreement</td>
</tr>
<tr>
<td>Whole study group</td>
<td>50</td>
<td>78%</td>
</tr>
<tr>
<td>High and intermediate density only</td>
<td>44</td>
<td>93%</td>
</tr>
<tr>
<td>Intermediate density only</td>
<td>26</td>
<td>65%</td>
</tr>
<tr>
<td>High and low density only</td>
<td>30</td>
<td>93%</td>
</tr>
</tbody>
</table>

* Selected using original density category assignment from FMMP mammography codes. \( ^b \) Agreement between density category, assigned using original FMMP mammography codes and density category selected by second radiologist. \( ^c \) Agreement between density category, assigned using FMMP mammography codes and assigned using FMMP codes selected by second radiologist. \( ^d \) Weighted kappa statistic.

When grouped by the calendar year of mammogram, inter-rater agreement was substantial for all year intervals \( [wK= 0.77, \ 0.69, \ 0.70 \text{ for } \leq 1995, \ 1996-2000, \ > 2000, \text{ respectively}] \). Only moderate agreement was observed when codes applied to the second radiologist reading were used to determine the density category (2nd-descriptors), compared to the original density category assignment \( [wK=0.55] \). In both comparisons (2nd-direct and 2nd-descriptors), the largest disagreement was observed between assignments to intermediate vs. low density category with the second radiologist being more likely to assign the low density category. Cross tabulation of the original breast density category assignment and that from 2nd-direct Q and 2nd-descriptors method indicated the exact agreement of 78\% (39 films) and 62\% (31 films), respectively (Table 2).

**Quantification of Breast Density from Digital Image**

Of 50 mammograms, 23 mammograms were scanned with resolution 1, 2 mammograms were scanned with resolution 2, 22 mammograms were scanned with resolution 3, and 3 mammograms were scanned with resolution 4. The middle triad was consistently denser (higher PAG value) compared to the lower and the upper triad across all original density categories. Cross tabulation of the breast density group assignment from digital data (adjusted for resolution) and the original breast density category indicated the exact agreement of 52\% (26 films) between two methods \( [wK=0.27] \). Better agreement was seen when the analysis was limited to the films originally categorized as high or intermediate density \( (Kappa = 0.36) \). None of the 24 originally high density mammograms was categorized as low density using the breast density estimation from the digital image; however, 8 out of 24 mammograms (33.33\%) were assigned intermediate density.

The results of the restricted analysis on resolution 1 group...
only (baseline, never adjusted) and the restricted analysis on the resolution 3 group only (using unadjusted PAG values) are presented in Table 3. When the analysis was restricted to the 39 mammograms with agreement in the density category assignment from the original mammography codes and the overall density assessment from the second radiologist’s reading, the best improvement was noticed for the subset of high and intermediate density groups. However, the kappa statistic remained within the fair to moderate agreement range (Table 3).

<table>
<thead>
<tr>
<th>Study group/subset</th>
<th>N</th>
<th>Kappa coefficient</th>
<th>[95% CI]</th>
<th>N</th>
<th>Kappa coefficient</th>
<th>[95% CI]</th>
<th>N</th>
<th>Kappa coefficient</th>
<th>[95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole study group</td>
<td>50</td>
<td>0.27 *</td>
<td>[0.06-0.47]</td>
<td>23</td>
<td>0.38 *</td>
<td>[0.09-0.38]</td>
<td>22</td>
<td>0.38 *</td>
<td>[0.07-0.69]</td>
</tr>
<tr>
<td>High and intermediate density only</td>
<td>44</td>
<td>0.36</td>
<td>[0.09-0.64]</td>
<td>20</td>
<td>0.39</td>
<td>[-0.01-0.80]</td>
<td>19</td>
<td>0.45</td>
<td>[0.04-0.85]</td>
</tr>
<tr>
<td>Intermediate and low density only</td>
<td>26</td>
<td>-0.03</td>
<td>[-0.35-0.28]</td>
<td>14</td>
<td>-0.09</td>
<td>[-0.53-0.36]</td>
<td>10</td>
<td>0.07</td>
<td>[-0.43-0.58]</td>
</tr>
<tr>
<td>High and low density only</td>
<td>30</td>
<td>0.24</td>
<td>[-0.08-0.57]</td>
<td>12</td>
<td>0.08</td>
<td>[-0.08-0.92]</td>
<td>15</td>
<td>0.33</td>
<td>[-0.16-0.83]</td>
</tr>
</tbody>
</table>

* Weighted kappa statistic.

| Figure 2. Examples of different density category assignment by two methods (original vs. digital). (Each mammogram represents a different woman.)

A. Low density mammogram as determined by FMMP codes assigned: A1-high density from digital images; A2-low density from digital images (agreement);
B. Intermediate density mammogram as determined by FMMP codes assigned: B1-high density from digital images; B2-intermediate density from digital images (agreement);
C. High density mammogram as determined by FMMP codes assigned: C1-intermediate density from digital images; C2-high density from digital images (agreement).

Discussion

This study was designed to find the best method to characterize mammographic breast density for a planned retrospective study. Our findings indicate that using the FMMP mammography codes for characterization of breast density is better than using information from digitized radiography films.

This study involved mammograms from a large cohort of women with multiple mammograms obtained over time. The
FMMP mammograms were performed over a 16 year period, which allowed a stratified selection of mammograms by both the density category and the calendar year of the mammogram. The retrospective study of breast density determinants will use only the extreme breast density phenotypes (low and high). However, we needed to include the intermediate density category in this study to detect disagreement in the assignment of original low or high density mammograms to the intermediate density category by either of the investigated methods. The number of mammograms in the study was limited by funding and time allocated for this pilot project. However, the significant results suggest that this sample size was adequate to detect the existing differences.

The original density assignment was based on the combined information from both craniocaudal (CC) and MLO views, and some information on density might have been lost by selecting only MLO views for this study. However, previous studies have reported high correlation between density levels estimated from CC and MLO views by either radiologists or computer-aided techniques (Martin et al., 2006, Jeffreys et al., 2003). The FMMP mammograms differed with respect to their technical specifications and quality. The use of four resolutions in this study was necessary to obtain the best information possible from the image. Comparison of the kappa statistics from the analyses of resolution 1 group only (never adjusted) and unadjusted PAG values for resolution 3 group only, with the results from all four resolutions combined suggests that the findings are similar, though still slightly affected by the adjustment.

Density categories from digital output were assigned by ranking all mammograms by the PAG values and assigning the number of mammograms to each category corresponding to the original distribution of the density categories in the sample. Other methods for assigning density categories from the digital output were also explored. We considered using overlap areas between 95% confidence intervals for the PAG values across three density groups to assign films with the PAG values within the overlap ranges into the other closest density category. However, this method could not be applied as the confidence intervals for intermediate (100.5-197.0) and low (101.4-191.9) density completely overlap the much more narrow confidence interval for high density (130.9-195.6), which did not allow us to clearly define PAG ranges for each of the density categories. We also considered using a cumulative distribution plot for PAG values to detect any natural cut points for separating images into the density categories, but there was no distinct point for change in slope that would define the intermediate and high density categories.

For all three original density categories, the images that were assigned the same density categories from the original radiologist’s reading and the digital output (concordant) were compared against the images that were assigned different categories by two methods (discordant). Concordant mammograms tended to have dense pixels more concentrated in certain areas of the image, while discordant mammograms consistently showed more diffuse distribution of dense pixels throughout the image (Figure 2). It is possible that using PAG does not take into account the distribution pattern of the dense pixels within the image.

Previous studies have noted that fatty breasts tend to display more diffuse breast tissue patterns (Klifa et al., 2004). In our study, the diffuse pattern made density misclassification by the digital output more likely and we expected to see a difference in BMI between women with concordant vs. discordant mammograms. For the high density category with the best agreement between two methods, there was no difference in BMI between women with concordant and discordant mammograms ($p=1.00$). There was a statistically significant difference in BMI between women with discordant mammograms in the original intermediate density category (23.7 and 26.4, respectively; $p=0.04$), but no significant difference was found for the original low density category, which might be a result of a small number of films. However, BMI in women with discordant mammograms was lower (19.4 and 23.0, respectively).

The middle triad was found to be denser compared to the lower and the upper triads across all three original density categories. These findings were consistent with previously reported results suggesting that the denser area corresponds to the central regions of the breast (Li et al., 2004).

Our findings on agreement between breast density estimation by different radiologists are consistent with the previously reported results (Ooms et al., 2007, Ciatto et al., 2005, Roubidoux et al., 2003, Berg et al., 2000, Jong et al., 1996). Previous studies demonstrated that the overall classification of density from digitalization of old mammography films tended to be good, but the breast density assignment from the digital images tended to be low compared with the expert (Muhimah & Zwiggelaar, 2006, Lee-Han et al., 1995, Jeffreys et al., 2003). Although studies suggested that computer-assisted techniques could be used for categorization of breast density, the classification of breast densities in those studies differs from the approach used to categorize breast density on the FMMP mammograms (Jeffreys et al., 2003, Lee-Han et al., 1995, Muhimah et al., 2006). The computer-assisted methods in previous studies were also different from the method used in our study (PAG values).

Based on these results, we will add additional codes to improve definition of the density phenotype. The results of this methods study are important for researchers conducting retrospective studies of breast density using radiography films. Our findings indicate that a dual reading of the mammograms by two radiologists specifically asked to assess the level of breast density provides a better estimation of breast density and may be much less expensive.

Acknowledgements

Support for this project was provided by the Fernald Medical Monitoring Program and the UC Research Council as a Graduate Student Summer Research Fellowship awarded to Yaghjyan.

References


